STAT 672 Statistical Learning & Data Analytics Spring 2018

George Mason University

Instructor: Martin Slawski

Homework 5

Pencil-and-paper part due Thursday, April 19th, in class. Code and relevant output has to be submitted via Blackboard. Please follow the following format for the filename of your submission: `Lastname_Firstname_HWx.zip`, where `x` needs to be substituted by the homework #.
Maximum possible points to earn: 18; 100% = 15 points, rest counts as bonus.

Feature selection and cross-validation

The goal of this homework is to make you aware of potential pitfalls in the use of cross-validation.

In class, cross-validation was introduced for the purpose of model selection. In addition, cross-validation can also be used for estimating the generalization error. The general idea is to use multiple hold-out test sets, instead of a single hold-out set. There are several ways of generating multiple hold-out sets such as repeated random subsampling, or, as mentioned, cross-validation. As opposed to random subsampling, cross-validation ensures that all data points appear at least once in a hold-out set.

Part I

a) Generate a n-by-d matrix $\mathbf{X}$ with i.i.d. $N(0,1)$-entries, with $n = 100$, $d = 10,000$, and a vector of labels $\mathbf{y} = (1, -1, 1, -1, \ldots, 1, -1)^\top \in \mathbb{R}^n$.

b) Select a subset S of the $10,000$ features by taking the top five features with largest magnitude among the statistics $\{|c_j|\}_{j=1}^d$, where

$$c_j := \frac{1}{n}\sum_{i=1}^n X_{ij} y_i.$$

c) Apply ten-fold cross-validation[1] with $(\mathbf{X}_S, \mathbf{y})$ and logistic regression (with intercept, without centering/scaling[2]), where $\mathbf{X}_S$ denotes the sub-matrix of $\mathbf{X}$ obtained by extracting the columns of $\mathbf{X}$ corresponding to S.

[1] Code it yourself. Do not use external functions.

[2] This does not play much of a role for the problem that I want to illustrate here.

d) What do you observe for the cross-validation error, and what would one expect under the data-generating process in a)? Explain where the discrepancy comes from.

Part II

You are given a data set from some biological experiment - the features represent physiological parameters of a group of individuals and the label is 1 if the individual has a certain disease and -1 otherwise. The biologists who have generated the data claim that they can predict the labels with around 80% accuracy, using least squares together with feature selection.

Your task is to check if the biologists have done a good job in their data analysis. Check their claim by implementing an exhaustive search for the best feature subset. As a classifier use least squares, that is the weight vector is given as

$$\widehat{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

where $\mathbf{X}$ is the feature matrix and $\mathbf{y}$ the vector of labels.
Classification of a new test instance x is done via $f(x) = \text{sign}(\langle \widehat{w}, x,) \rangle$. As error measure use the classification error,

$$L(Y_i, f(X_i)) = \frac{1}{2} |Y_i - \text{sign}(\langle \widehat{w}, X_i \rangle)|.$$

There are $d = 15$ features in the training data. Use 5-fold cross-validation (use the ordering of the data as it is provided*) for the linear least squares classifier in order to determine the best feature subset among all possible $2^{15} - 1 = 32,767$ possible feature subsets** by minimizing the 5-fold cross validation error on the training data `(Xtrain,Ytrain)`.

a) Report the best feature subset(s) and its/their 5-fold cross-validation error (**written on paper**).

b) Use the whole training data of the best feature subset(s) obtained in a) to learn the final classifier. In the meantime the biologists have obtained new data. Use this data, given as `(Xtest,Ytest)` to evaluate the performance of the classifier(s) and report it. Do you have an idea why the cross-validation error obtained in a) and the just computed test error are so different? Has this to do with the classifier or the feature selection? What will you tell the biologists?

Hints:

* the indices for the five folds need to be taken as $\{1, 2, \ldots, n/5\}, \{n/5 + 1, \ldots, 2n/5\}, \ldots, \{4n/5 + 1, \ldots, n\}$.

** Use existing code to generate all possible feature combinations (unless you really want to do it yourself, but then make sure your code works). In `Python`, I recommend using `itertools.combinations`.

c) Explain how randomly permuting the labels `Ytrain` could have been used to detect the issue in b) without the need of an additional test set.

d) Implement the idea in c) for 20 random permutations***. For each of those random permutations, re-run a) with `Ytrain` being replaced by its permuted version, and save the minimum cross-validation error over all feature subsets. How do the minimum errors obtained in this way compare to the one in a)?

*** Depending on your machine, this may take between five and ten minutes of time.

Point distribution:

Ic)	Id)	IIa)	IIb)	IIc)	IId)
3	3	5	3	2	2